NOTES 04: CATEGORICAL VARIABLES

Stat 120 | Fall 2025

Prof Amanda Luby

1 Categorical Variables

Categorical variables are best summarized with a **frequency table** and visualized using a **barplot**. When we want to summarize a categorical variable with a single number, we often use a **proportion**.

> Proportion

When we have two categorical variables, we often use a **two-way table** to summarize them at the same time (also called the **joint distribution**). We might also care about the **marginal distribution** (the margins) or **conditional distribution** (a specific row/column).

Example: Below is the two-way table for our class representing the answers to "Have you taken a CS class before?" and whether the "Environmental Issues" interest box was checked.

	Week 1	Week 7
Yes (Got Sleep)	10	4
No (Not enough sleep)	6	9

a. What is the marginal distribution of *Sleep*?

b. What is the conditional distribution of *Week* among those who *did not get enough sleep*?

c. What is the conditional distribution of *Sleep* among those who *were surveyed in Week 1*?

d. What is the proportion of students who were surveyed in Week 1?

e. Does *sleep* appear to be independent of *week*?

2 Quantitative Variables

Quantitative variables are best visualized with a **histogram** or **dotplot** (depending on sample size)

When describing quantitative variables, we typically care most about the **shape** and **center**. When we want to summarize a quantitative variable with a single number, we often choose the **mean**, **median**, or **mode**.

Skewed Right	Symmetric	Skewed Left

.

There are various ways to describe the center of the distribution. The three most common are:

Mean

Median

Mode